Deep Learning

880663-M-6

Assignment

Using Deep Learning to Perform Multi-Class Classification on the Lung and Colon Cancer Histopathological Image Dataset (LC25000)

Report by:

Sevda Georgieva (2121325)

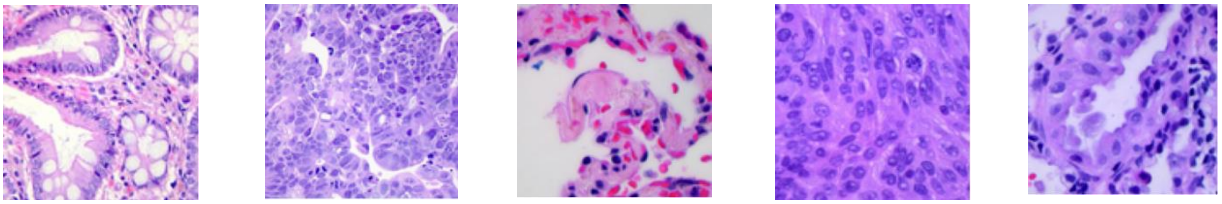
March 2024

Exploratory Analyses

The EDA analyses concerned a random visualisation of 15 images from the dataset. This was done to get aquainted with the image quality and characteristics of each cancer class (see Fig. 1).

Figure 1

Sample Image from each Class



Colon Benign Tissue | Colon Adenocarcinoma | Lung Squamous Cell Carcinoma | Lung Squamous Cell Carcinoma | Lung Adenocarcinoma

Additionally, I used barcharts to visualise the data distribution between the five classes and observe for possible class imbalance. The barcharts indicated an even distribution with 5000 images per class (see Fig. 2).

Figure 2

Class Distribution

(a) (b) (c) (d) (e)

(a) Colon Benign Tissue

(b) Colon Adenocarcinoma

(c) Lung Benign Tissue

(d) Lung Squamous Cell Carcinoma

(e) Lung Adenocarcinoma

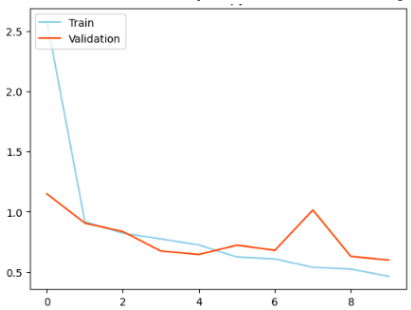
Results of the Baseline Model

Loss and Accuracy

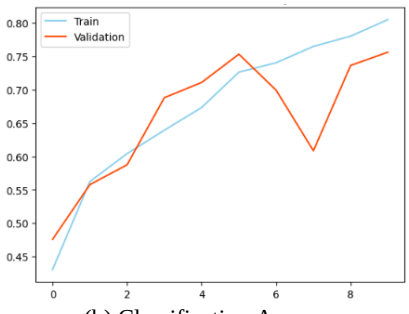
Figure 3a shows that the training loss diminishes with newer epochs, while the validation loss increases. Figure 3b shows that while training accuracy steadily increases, validation accuracy varies and is lower. The observations from both figures suggest that the model is overfitting to the training data, capturing details and noise that do not generalize to the validation data.

Figure 3

Cross-Entropy Loss and Classification Accuracy of Baseline Model



(a) Cross-Entropy Loss



(b) Classification Accuracy

Confusion matrix

The confusion matrix for the validation set shows a failure to discriminate between certain classes. For example, ‘Colon Benign Tissue’ is misclassified as ‘Colon Adenocarcinoma’ and vice versa. A parallel misclassification pattern is observed between ‘Lung Squamous Cell Carcinoma’ and ‘Lung Adenocarcinoma’ (see Fig. 4a). ‘Lung Benign Tissue’ is the only class with good classification, which I assume is due to its distinctive red patterns in the images (see Fig. 1). The

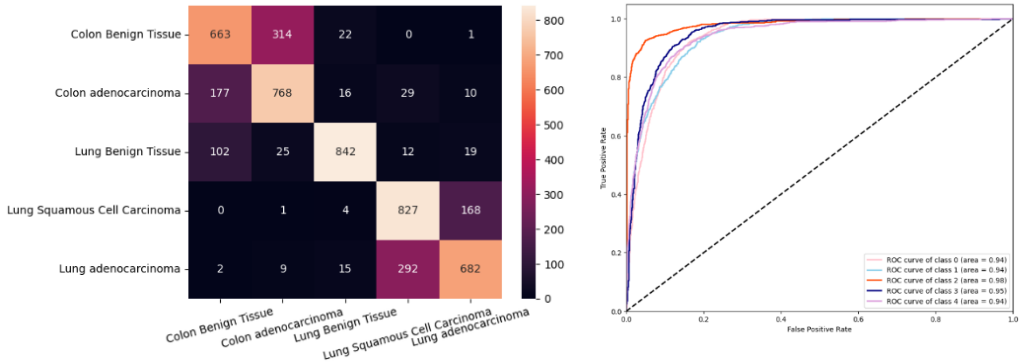
relevance of false positives and false negatives in comparison between ‘Colon Benign Tissue’ and ‘Colon Adenocarcinoma’ as well as between ‘Lung Squamous Cell Carcinoma’ and ‘Lung Adenocarcinoma’ suggests that while the model can discriminate whether cancer is in the lung or colon, it fails to differentiate between specific cancer classes.

ROC curves

The ROC curve for the validation set shows high to medium model classification performance with all ROC AUC ObV being over 0.90 (see Fig. 4b). Though this indicates effective class separation, the confusion matrix above suggests pair-wise classification challenges. Thus, while the baseline model is generally effective in differentiating one class from all other classes, improvements are needed for the model‘s performance in distinguishing between specific classes.

Figure 4

Confusion Matrix and ROC curve of Baseline Model’s validation set



(a) (b)

Performance measures

The performance metrics of the model demonstrate similar results for the validation set, with precision, recall, F1-score, and accuracy spanning a wide range, from approximately 66% to 94% across the five classes. This broad value spread indicates a varied level of predictive reliability for different classes. The values for each metric correspond to the output of the confusion matrices i.e., the model performance is good only for 'Lung Benign Tissue', while other classes are misclassified, leading to false positives and false negatives (see Table 1).

Table 1:

Baseline model performance on the validation set

	Validation set			
	Precision[%]	Recall[%]	F1[%]	Accuracy[%]
Colon Benign Tissue	0.70	0.66	0.68	
Colon Adenocarcinoma	0.69	0.77	0.73	
Lung Benign Tissue	0.94	0.84	0.89	
Lung Squamous Cell Carcinoma	0.71	0.83	0.77	
Lung Adenocarcinoma	0.78	0.68	0.73	
				0.76

Improved (Fine-tuned) Model

Experimentation

My strategy for enhancement of the baseline model was to reduce overfitting, improve generalization, and improve the model differentiation between specific classes. I engaged in an experimental approach, where I changed parameters one at a time to properly evaluate their partial effect on model performance. The selection of parameters for experimentation was firmly based on literature, however, the final parameter values were determined empirically.

I started by exploring different optimizers as they can significantly impact the performance of the baseline model without adding complexity. Considering the complex nature of histological image data, my selection concerned only adaptive optimizers as they allow for faster convergence and an

overall stable training process (Goodfellow et al., 2016). I used RMSprop, Nadam, and Adamax all of which are used in image classification tasks (Ahmmed et al., 2023), (Kandel et al., 2020). Each optimizer was tested with a learning rate ranging from 0,0001, to 0.01. After comparing the baseline model performance with each optimizer, Adamax was the superior choice, giving the highest accuracy and lowest loss values of 0.87 and 0.43 respectively. I retained the learning rate of 0.001 as neither its increasing nor decreasing resulted in significant performance gains.

To avoid the 'dying ReLU' problem of the ReLU function, I used a leaky ReLU with a standard alpha of 0.01 to mitigate the ReLU limitation of zero output for negative input (Goodfellow et al., 2016). The use of leaky ReLU increased the accuracy to 0.90 and decreased loss to 0.41.

The relatively simplistic architecture of the baseline model is likely one reason for the model's difficulty in distinguishing between certain classes. Increasing the depth and width of the baseline model would enable it to capture a hierarchy of features, improving image classification (Zeiler & Fergus, 2014). A standard approach is to follow a pyramid structure, which starts with fewer neurons in the initial layers, progressively increasing their amount with each new layer (Ullah & Petrosino, 2016). This approach allows for a gradual and systematic extraction of features, which can lead to improved model generalization. I explored different pyramid structures, I chose the second best performing model to balance between model performance and computational efficiency. This model gave accuracy of 0.96 and loss of 0.11. Additionally, I implemented early stopping, setting the epoch range to 50 and patience to 10, to find the optimal number of epochs for training (Goodfellow et al., 2016). Early stopping indicated 12 epochs to be ideal as further training did not improve model performance. The increase in structure complexity and epochs lead to 0.98 and 0.13 for accuracy and loss respectively.

While model complexity improved model performance, overfit remained (see Jupyter notebook 'Model Fit of Final Model'). Thus, I investigated two regularization techniques - L2 and dropout rate. I compared model fit both by using the techniques individually and together. The combined approach aimed to prevent the learning of overly complex patterns, indicative of noise, and encourage the development of more robust features by introducing randomness (Goodfellow et al., 2016). The L2 values I tested ranged from 0.0001 to 0.01, with 0.001 giving the best performance. The tested dropout values ranged from 0.15 to 0.50, with 0.20, 0.25, and 0.25 in blocks two, three, and four respectively giving the best performance. The combined model used the best performing L2 and dropout. Only dropout was chosen for the final model because it gave the lowest loss (0.07) and highest accuracy (0.98) and also led to a good fit between the training and validation set (see Jupyter notebook – 'Model Fit Comparison').

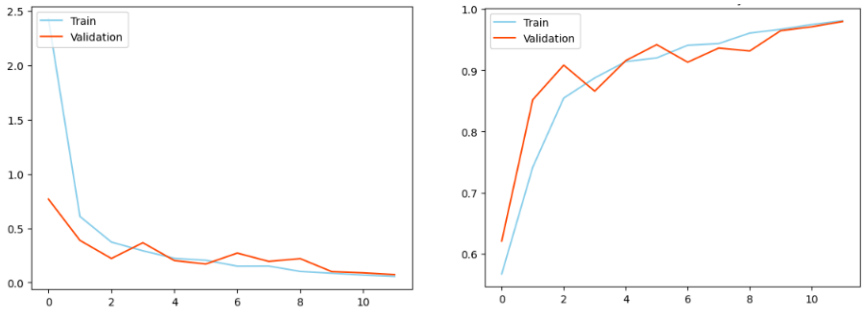
Results of the Enhanced Model

Loss and Accuracy

Figure 5a shows an improvement in the enhanced model over the baseline where the enhanced model demonstrates a significant improvement with both training and validation losses decreasing in tandem. This implies that the model is learning generalized patterns rather than noise. Figure 5b reveals further advancements with the enhanced model exhibiting a continuously high training accuracy that is closely matched by the validation accuracy. The accuracy remains relatively stable across epochs without a significant downward trend. Both figures indicate that the enhanced model maintains its predictive performance on unseen data, solving the overfitting issue observed in the baseline model.

Figure 5

Cross-Entropy Loss and Classification Accuracy of Enhanced Model



(a) Cross-Entropy Loss (b) Classification Accuracy

Confusion Matrix

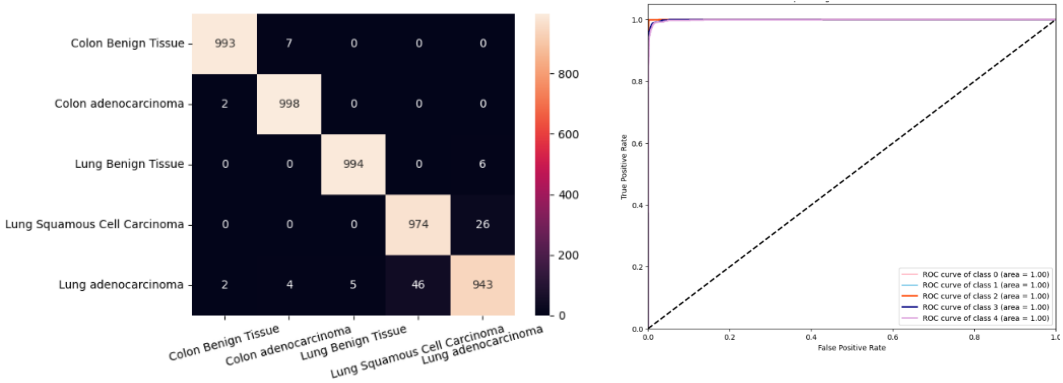
The confusion matrix for the enhanced model demonstrates a substantial performance improvement over the baseline model. Particularly it shows a considerable reduction in the misclassifications between 'Colon Benign Tissue' and 'Colon Adenocarcinoma', evident from the higher true positive rates and fewer off-diagonal elements in the confusion matrix for the test set. This suggests that the enhanced model is better at capturing subtle differences between the classes (see Fig. 6a). However, while the model has improved its predictions between 'Lung Squamous Cell Carcinoma' and 'Lung Adenocarcinoma', the prevalence of false positives and negatives suggests that the differences between the two classes is still too subtle for the enchanced model. This argues the need for further optimization or use of more complex CNN architectures.

ROC Curve

The ROC curves for the enhanced and baseline model show a distinction in their performance capabilities. The curves of the enhanced model (Fig. 6b) show an exceptional case of perfect classification with success of achieving a 100% positive rate in class separation. This performance is better than that of the baseline model which demonstrated good but not perfect classification performance. Combined with the confusion matrix, these results suggest that the enhanced model not only improved in general class separation but also accurately discriminates between specific classes.

Figure 6

Confusion Matrices of the Baseline Model for Test Set



(a) (b)

Performance Measures

The comparison between the baseline and enhanced model reveals a dramatic improvement in the latter's performance metrics. The enhanced model presents nearly perfect performance metrics which range between 0.95 to 1.00 for all classes (see Table 2). This means the enhanced model correctly classifies almost all instances classified as positive (i.e., precision) and accurately identifies almost all positive cases (i.e., recall). Additionally, the F1-score, which balances the two prior metrics, further confirms that the enhanced model provides a balance in identifying positive cases and minimizing the false positive ones. This elevates the overall accuracy to 0.98, suggesting a highly improved model able to generalize and accurately reflect real-world data.

Table 2:
Enhanced model performance

	Test set			
	Precision[%]	Recall[%]	F1[%]	Accuracy[%]
Colon Benign Tissue	1.00	0.99	0.99	
Colon Adenocarcinoma	0.99	1.00	0.99	
Lung Benign Tissue	0.99	0.99	0.99	
Lung Squamous Cell Carcinoma	0.95	0.97	0.96	
Lung Adenocarcinoma	0.97	0.95	0.96	
				0.98

Transfer Learning Model and Its Results

I chose ResNet-50 as its characteristics cover the issues observed with the baseline model. Specifically, the model's deep architecture makes it a good choice for complex tasks. For example, (Sahaai et al., 2022) employed this model in a multiclass classification of brain tumor using MRI data. The model's accuracy was 95.3%. The results suggest that using ResNet-50 as my transfer learning model will detect the subtle differences between the five classes and reduce the misclassification instances observed in the baseline model.

This model did not undergo any optimization to reduce computational load and introduce a baseline for the transfer learning model. The number and density of the flattened layers mimicked that of the enhanced model. This aimed to maintain consistent feature extraction and allow for a direct performance comparison between the transfer and enhanced models (refer to Jupyter file 'ResNet-50 Model Summary').

Model fit

While small, the model shows a degree of overfit due to discrepancies between the training and validation loss and accuracy (refer to Jupyter file 'Loss and Accuracy of ResNet-50'). This is likely due to the increased model complaxity and lack of regularisation functions which enable the learning of noise.

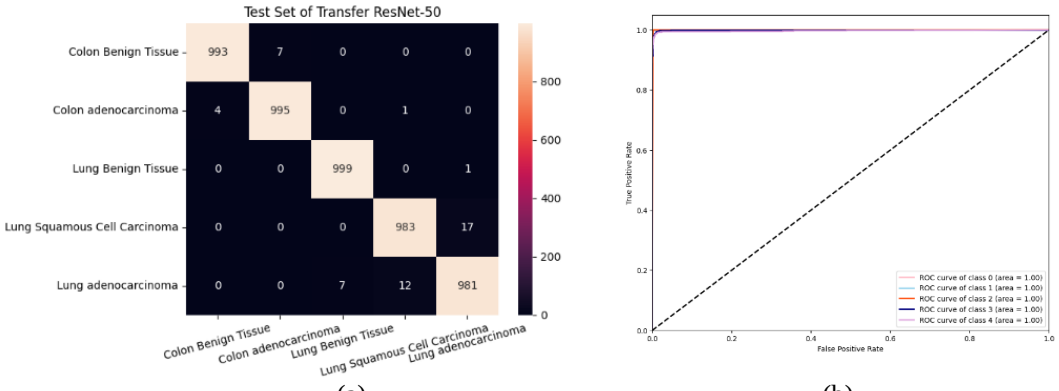
Confusion Matrix

The confusion matrix for the transfer learning model show improved performance compared to the enchanted model in differentiating between 'Lung Squamous Cell Carcinoma' and 'Lung Adenocarcinoma'. The high number of true positive rates and few off-diagonal elements suggest that the model successfully distinguishes between closely related classes and produces fewer misclassifications (see Fig. 7a).

ROC Curve

The transfer learning and enhanced models are identical in their ROC curves, where both models maintain a high true positive rate of 1 (see Fig. 7b).

Figure 7
Confusion Matrix (a) and ROC curve (b) for the ResNet-50 on the Test set



(a) (b)

Performance Measures

The transfer learning displays high-performance metrics, with values from 0.98 to 1.00 for all classes (see Table 3). Similar to the confusion matrics the values for ‘Lung Squamous Cell Carcinoma’ and ‘Lung Adenocarcinoma’ arehigher compared to the enchanced model. The transfer learning model exhibits excellent precision and recall, leading to a high F1-score showing its ability to accurately identify positive cases while minimizing false positives. Consequently, this results in an overall model accuracy of 0.99, indicating a significantly refined model.

Table 3:

Performance Metrics of the ResNet-50 on the test set

	Test set			
	Precision[%]	Recall[%]	F1[%]	Accuracy[%]
Colon Benign Tissue	1.00	1.00	1.00	
Colon Adenocarcinoma	0.99	1.00	1.00	
Lung Benign Tissue	1.00	1.00	1.00	
Lung Squamous Cell Carcinoma	0.99	0.98	0.98	
Lung Adenocarcinoma	0.98	0.98	0.98	
				0.99

Discussion

One limitation of this paper is the lack of explored pre-processing techniques which could have been useful given the low image resolution of 120x120. (Chehade et al., 2022) used visual analyses such as Gaussian blur and unsharp masking on the same dataset which reduced noise and sharpened image features. The best performing hybrid model which was XGBoost, achieved an accuracy of 99% and a F1-score of 98.8%. Additionally, data augmentation techniques such as horizontal and vertical rotations and flips increase the diversity of the training dataset, which make the model more robust and lead to better generaliastion (Garg&Garg, 2020). Grag and Grag (2020), used these augumentation techniques in a ResNet-50 transfer learning model which led to 100% precision, recall, F1-score, and accuracy for lung cancer and colon cancer identification. The inclusion of such pre-processing techniques could also improve the performance of my enchanced model, which produced missclassifications between ‘Lung Squamous Cell Carcinoma’ and ‘Lung Adenocarcinoma’. The missclassification could also be improve by increasing model complexity as shown by the ResNet-50 which included 50 layers and increased true positives between the two classes.

Furthermore, beyond the ResNet-50 model, Greg and Grag (2020) also compared the performance of other transfer leanring models such as VGG16, InceptionV3, InceptionResNetV2, MobileNet, Xception, NADNetMobile, and DenseNet169, where precision, recall, F1-score and accuracy range from 96% to 100%. This suggests that more complex architectures, in general will outperform my enchanced model. Yet, excess model complexity could lead to overfitting, high computational load, and lack of interpretability (Goodfellow et al., 2016). An alternative could be exploration of hybrid models which use features extracted from CNN’s and classifiers from machine learning (ML), offering a balance between deep learning’s (DL) complexity and relative simplicity of ML models. For example, one study used this approach where features were extracted by DenseNet-121 and classified by a Random Forest. They classified colon cancer histological images, where the model achieved accuracy of 98.60%, precision of 98.63%, and a recall of 98.60%. These results, along with the results of Chehade et al., (2020) mentioned above, are on par with traditional DL models and suggest the potential of hybrid models to provide valuable insights for accurately classifying the five cancers in the LC2500 dataset.

The discussed literature advises future researchers to consider the above strategies to overcome the limitations of the enchanced and transfer learning model, in an effort to advance the accuracy and reliability of cancer diagnosis through histopathological image analysis.

References

Ahmmed, S., Podder, P., Mondal, M., Rahman, S., Kannan, S., Hasan, M., Rohan, A., & Prosvirin, A. (2023). Enhancing brain tumor classification with transfer learning across multiple classes: An in-depth analysis. *BioMedInformatics*, *3*(4), 1124–1144. https://doi.org/10.3390/biomedinformatics3040068

Chehade, A. H., Abdallah, N., Marion, J.-M., Oueidat, M., & Chauvet, P. (2022). *Lung and Colon Cancer Classification Using Medical Imaging : A Feature Engineering Approach*. https://doi.org/10.21203/rs.3.rs-1211832/v1

Garg, S., & Garg, S. (2020). Prediction of lung and colon cancer through analysis of histopathological images by utilizing pre-trained CNN models with visualization of class activation and saliency maps. *2020 3rd Artificial Intelligence and Cloud Computing Conference*. https://doi.org/10.1145/3442536.3442543

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Kandel, I., Castelli, M., & Popovič, A. (2020). Comparative study of first order optimizers for image classification using convolutional neural networks on histopathology images. *Journal of Imaging*, *6*(9), 92. https://doi.org/10.3390/jimaging6090092

Sahaai, M. B., Jothilakshmi, G. R., Ravikumar, D., Prasath, R., & Singh, S. (2022). ResNet-50 based deep neural network using transfer learning for Brain tumor classification. *INTERNATIONAL CONFERENCE ON RECENT INNOVATIONS IN SCIENCE AND TECHNOLOGY (RIST 2021)*. https://doi.org/10.1063/5.0082328

Ullah, I., & Petrosino, A. (2016). About pyramid structure in Convolutional Neural Networks. *2016 International Joint Conference on Neural Networks (IJCNN)*. https://doi.org/10.1109/ijcnn.2016.7727350

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding Convolutional Networks. *Computer Vision – ECCV 2014*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53